Crowdsourcing Microdata for Cost-Effective and Reliable Lexicography

Martin Benjamin
École Polytechnique Fédérale de Lausanne
martin.benjamin@epfl.ch

Abstract

Lexicography has long faced the challenge of having too few specialists to document too many words in too many languages with too many linguistic features. Great dictionaries are invariably the product of many person-years of labor, whether the lifetime work of an individual or the lengthy collaboration of a team. Is it possible to use public contributions to vastly reduce the time and cost of producing a dictionary while ensuring high quality? Crowdsourcing, often seen as the solution for large-scale data acquisition or analysis, is fraught with problems in the context of lexicography. Language is not binary, so there may be no one right answer to say that a word "means" a particular definition, or that a word in one language "is" the same as a particular translation term. People may misinterpret instructions or misread terms or make typographical or conceptual errors. Some crowd members intentionally add bad data. Without a payment system, incentives for participation are slim; micro-payments introduce the incentive to maximize income over quality.

Our project introduces a public interface that breaks lexicographic data collection into targeted microtasks, within a stimulating game environment on Facebook, phones, and the web. Players earn points for answers that win consensus. Validation is achieved by redundancy, while malicious users are detected through persistent deviations. Data can be collected for any language, in an integrated multilingual framework focused on the serial production of monolingual dictionaries linked at the concept level. Questions are sequential, first eliciting a lemma, then a definition, then other information, according to a prioritized concept list. The method can also be used to merge existing data sets. Intensive trials are currently underway in Vietnamese, with the inclusion of additional Asian languages an explicit objective.

Keywords: Crowdsourcing, Gamification, Distributed lexicography, Multilingual, Quality assurance

1. Introduction.

Kamusi GOLD is the Global Online Living Dictionary, with the goal of producing a multilingual dictionary with comprehensive data for "every word in every language". Given the size of each language, the large amount of information and nuance that can be associated with each term, the number of languages and variants, and the near impossibility of extensively engaging informants for many smaller languages, the goal of "every word" is axiomatically out of reach. However, "every word" sets a target for system design: a structure that can accommodate the full range of linguistic data associated with each term (Benjamin 2014 b and c), and a set of procedures that can elicit and process data with speakers of any language (Benjamin 2014 a, Benjamin and Radetzky 2014 a and b). In this paper, we describe the public interface that breaks

lexicographic data collection into targeted microtasks. Tasks build upon each other, including the creation of new data and the validation of others' contributions. Questions are presented within stimulating game environments (McGonigal 2011, Hamari and Koivisto 2013) on Facebook, mobile devices, and the web. The ultimate goal is a resource that can give people and technology services a wide assortment of useful information for any written or spoken term. Each language should have a monolingual dictionary of all the concepts it produces, with definitions in its own language, extended information such as word forms, and links to close equivalent ways of expressing the idea in other languages. Our premise is that most of the data we are seeking cannot or will not be produced in a codified form by relying on traditional methods of field or desk lexicography. Through well-regulated public engagement, the methods described in this paper can collect data systematically for numerous languages that would otherwise be out of reach for detailed lexicographical investigation.

2. Challenges Addressed

2.1 Great dictionaries take time. Individuals often compile dictionaries over decades, and teams may not be much faster. Acquiring and arranging lexicographical data is a difficult task that requires serious oversight from knowledgeable individuals. Rare is the case where lexicography pays the rent, and rare is the case where a language enthusiast can see a serious dictionary through to completion; because of limitations of time and person power, great dictionaries for most languages have never been written. The Kamusi Project seeks to change that, through a combination of systems for language experts to compile dictionaries using a central lexicographical resource, computational techniques for harvesting existing data, and the methods discussed in this paper for users to contribute to the growth of resources for their own languages. The crowd will still need many person-years to achieve quality results, but the metered time can be condensed through an expanded number of participants versus traditional lexicography (Brabham 2011). With careful control of data acquisition and review, we contend that public engagement will result in quality lexicographical data for many languages that otherwise would not or could not exist.

2.2. Crowdsourcing offers the potential of rapid data collection at low cost, but it is also fraught with dangers (Saxton et al. 2013). Crowdsourcing with micropayments through services such as Mechanical Turk introduces a host of incentives for participants to cheat (Kittur et al. 2008, Hirth et al. 2013) that we would not want to expend resources in combatting, were our budget even to allow. Absent a financial incentive, many people with great intentions may nevertheless provide poor information. They may have a faulty understanding of a question, or an inadequate understanding of the parameters of a satisfactory answer. They may have poor spelling or poor grammar. Additionally, users might provide wrong information for malicious reasons. Many crowdsourcing projects surmount such problems by greatly limiting the tasks that contributors can perform (Hsueh et al. 2009), with limited goals such as identifying the features of photos. Other projects ferret out bad information by relying on users to find problems, whether the Wiki method of trusting that mistakes will be found and fixed, or Amazon asking users if they find a review helpful. The methods we describe are designed to learn from each contributor to the maximum of their capabilities, while constraining the information to

the format needed for consistent data, and ensuring that bad data does not get published to the project.

2.3. In the context of language, crowdsourcing has peculiar advantages as well as posing particular problems. People are passionate about their languages in a way they are not about their reviews on Trip Advisor or their votes on Slashdot. Their passion may become dogmatic, to the extent that one speaker may try to overwrite contributions from someone who speaks a variant form. Even answers that are essentially the same may adhere to different orthographies, which the methods discussed herein do not address other than to pass conflict to a separate stage of manual review. Yet, important variations in language can be revealed by the crowd in a way not accessible to most lexicography projects. Both the numbers of potential participants and the ability for people to contribute along the entire geographic range across which a language might vary open public lexicography to a never-before-seen breadth of speaker input. For most crowd projects, participants can offer knowledgeable opinions, whether confirmation that an image from a telescope has the characteristics of a galaxy, or a fully-researched article on Wikipedia (Schenk and Guittard 2009). With language, however, a native speaker is an inherently authoritative, if problematic, source. The challenge for the project is to transfer this authoritative knowledge to a codified, reliable form.

3. Methods for Data Elicitation and Validation

This section describes the systems we have designed for eliciting valid lexical data from members of the public. The system is under continuous development, with frequent modifications, so the description at the time of writing might not exactly correspond to the implementation at the time of reading. Each of the modes has the same general facets: tasks should be simple to understand, simple to complete, and acquire validity through consensus. Some of the modes can be activated as-is for any language, while other modes need to be configured for the particular needs of a language (e.g., eliciting the dual form of nouns in Arabic). All modes would ideally have their controls and explanations properly localized for their own language, though this will require fairly extensive direct interaction with speakers to implement.

3.1 Translation Terms. Our first mode is geared to producing a set of parallel concepts across languages, in order to establish a baseline vocabulary for each. If we do not already have digitized data available for a language, which would be treated in the Alignment mode, this will be the first set of information that enters the system for the language. Although the project aspires to go well beyond crude mot à mot translations, this mode gives a foundation on which more complex data can be assembled.

In Translation mode, the participant is shown a word and definition in English, and asked for the word they would use to express that concept in their language. Importantly, we do not ask questions such as, “What is ‘pen’ in your language”, but rather, “Pen: a handheld writing implement with ink” or “Pen: an enclosure for animals”. The mode serves terms from an imperfect priority list based on factors such as English corpus frequency (Benjamin 2013). When Term X has multiple senses, we present the sense that appears first arbitrarily in our dataset, then move on through senses of the next ten terms on the

priority list, then cycle back to the next unqueried sense of Term X, so that the participant is not inundated with being asked about multiple senses of the same key term in rapid succession. When someone skips a question, we anonymously store that information so that we can learn which senses to prioritize and which to demote; e.g., all languages might have a way to express "run" as opposed to "walk", but not a "run" in baseball. When a threshold number of people submit the same term, we consider the item valid, pass it to the next mode, and publish it on Kamusi as a preliminary entry.

3.2. Synonyms. Although Translation mode can elicit the most popular term for a given concept, it also has the potential to bring in ambiguous information that needs a different mode for review. Were 10 Swahili speakers asked to translate "car", six might propose "motokaa", three might suggest "gari", and one could input "motocaa". Both "motokaa" and "gari" are valid answers, while "motocaa" is an error. In Synonyms mode, we are able to compare a validated term for which we have a definition with another term for which we have reason to suspect equivalence, either from original submissions via Translation mode or from the Open Multilingual Wordnet (http://compling.hss.ntu.edu.sg/omw/). We show the validated term and its definition, and ask whether the comparison term is another term with the same meaning, an alternate spelling of the same term, or an error. In the example, "gari" would quickly reach consensus and "motocaa" eliminated.

3.3. Word forms. Kamusi approaches morphological forms from two directions: cataloguing each form, or producing them algorithmically. When a part of speech has a small number of forms within a language, such as the five inflections of an English verb (see, sees, saw, seen, seeing), setting up procedures to elicit the forms is straightforward. In cases where the task is unambiguous, we can present the crowd with the lemma and its definition and ask for the desired subsidiary form, e.g., what is the plural for "car", a vehicle with four wheels?

This task has particular dangers. First, non-specialists are unlikely to be able to answer a question such as, "What is the past participle for 'see'"? Our current solution is to set up the Forms mode on a case-by-case basis, in consultation with language specialists; if we cannot find a way to phrase a question to get the proper form from the public, we will defer that item until we are able to assign the task to a trusted user. Second, we cannot assume that all senses of a lemma have the same inflected forms, e.g., the plural of the insect "louse" is "lice", while the plural of a human "louse" is "louses". Our solution is to ask for Forms for only a single sense, and then ask for confirmation that additional senses conform to the same spelling pattern – a low-threshold microtask. Third, alternate spellings are possible, such as "spelled" and "spelt". To solve this, we place competing spellings in the same type of verify/reject loop that we designed for Synonyms.

Certain parts of speech in certain languages have too many forms to tackle with iterative public input. The twenty or so forms that an adjective could have in many Bantu languages reaches the outer edge of what can be sought in Forms mode. Verb conjugation tables in Romance languages, which can hover near 100 forms, are best filled in by trusted users; a future tool will populate such tables. For agglutinative languages with

highly predictable forms, such as the nearly one billion possibilities for the Kinyarwanda verb, the only solution is to work with specialists on algorithms to parse the extended forms.

3.4. Definitions. “Dictionaries” for most languages are generally compilations of bilingual word pairs and selected extended information, with the assumption that readers have an a priori grasp on the underlying concept. For an almost random example, dictionaries for many languages pair a term in that language to English “bull”, with the typical first guess that the entry references a male cow. However, most good English monolingual dictionaries list more than a dozen other senses for “bull” as well. Our premise is that each term should be defined in its own language, creating monolingual references that link to the expressions of similar concepts across languages. Members of the public can provide these definitions, but writing a definition is a difficult task. The Definition mode is designed to work toward agreed definitions that could pass the muster of a trained lexicographer, with the acknowledgement that initial data might be less than perfect.

The starting point for the Definitions mode is the less-than-perfect set of ~200,000 English terms from the Princeton WordNet (PWN: http://wordnet.princeton.edu/). In the English version, we present an English term, such as “elevator car”, and the PWN working definition, in this case, “where passengers ride up and down”, and ask whether the given definition is good as is or if the reader can write a better definition. The reader also has the option to skip any entry, or to choose a definition submitted by an earlier participant. Within game play, a player receives substantial points for writing the winning definition, and one point for upvoting the eventual winning entry. Once a definition has passed the consensus threshold, it is published in Kamusi. However, Kamusi entries will have the option for readers to challenge the published definition, putting it back into consideration within the game setting.

For other languages, the Definition mode can be activated once a translation equivalent has been validated. We present the term, and show its English equivalent and definition (or PWN working definition). Showing the English definition unfortunately tends to steer the result toward the English concept rather than the nuance of the indigenous term, but is necessary for two reasons. First, it is absolutely essential that to align the correct senses across languages – to grab the right “bull” by the horns. Second, since most contributors do not arrive knowing how to write a good dictionary entry, the English definition will provide guidance that will give them a launching point, with the real potential that subsequent readers will have the incentive to improve weak definitions.

3.5. Examples. For languages that have a digital footprint, the written corpus can provide a trove of material for demonstrating the meanings conveyed in writing. The sense of a term cannot, however, be gleaned automatically, to a level of confidence one would inscribe in a dictionary. Expending enormous computational firepower to distinguish a raging bull from a bull market from a pile of bull would still result in only a statistical guess about whether a particular sentence pertains to a particular sense. The Examples mode combs through corpus sources, finds sentences that match one of the forms of a

term, and presents several of those sentences to reviewers. The participants' task is to select the lines that they feel exemplify a sense.

In Examples mode, a sentence must reach a high threshold to be accepted, but a lower one to be rejected. Once a few reviewers look at a sentence and decide it is not helpful, project time is better spent returning to the corpus and selecting another candidate for evaluation (Vaish et al. 2014). Finding a new example is cheap, whereas over-evaluating the same poor sentence costs time and goodwill. When set as a game, players receive points for all sentences where they are on the consensus side, either for selecting a good sentence, or for not selecting ones that others also agree are inadequate.

The initial version of Examples uses Twitter to harvest English text and the Helsinki Corpus of Swahili (HCS) for Swahili. We display the relevant term and definition, and for Swahili also the English translation and definition if available, and ask the user to click on all the results that are "excellent" examples of the sense in question. When we have three good examples for a sense, we publish them and remove the sense from rotation. For English, we use the Twitter API to find a number of English tweets that contain the term of interest to us, filtering for offensive terms. Tweets are messy text, with slang, misspellings, shortenings, hashtags, call signs, links, and languages mixing, so a large proportion of English tweets are unsuitable. Furthermore, experience teaches us that terms with many senses are too diffuse for this method. We therefore throttle this mode to terms with three or fewer senses in our database. A future generation of Examples will allow users to match candidate sentences to the list of available definitions.

For Swahili examples, we query HCS, which contains a large body of text from Swahili newspapers and books. HCS has parsed and tagged every term for its lemmatic form, which usually corresponds to the headword field in the Kamusi database. We thus already know that a sentence with a form such as "alipokitupa" is a candidate to exemplify "tupa", leaving our users the task of deciding if it exemplifies a particular sense of "tupa".

Future development of the Examples mode will build on both Twitter and the experience of mining HCS. The Indigenous Tweets project identifies people who usually tweet in a minority language. We will use that information to restrict queries for a language to its known producers. We will also apply the technique to more conventional text corpora for other languages, when the data is freely available.

3.6. Alignment. The biggest lexicographical barrier to uniting existing data in pursuit of a universal multilingual dictionary is the task of aligning concepts across languages. There is no computational way to know which English sense to choose for a term matched to "bull" from Vietnamese to English, for example, nor from Indonesian to English, and using the English spelling connection to link Vietnamese to Indonesian would multiply the risks of error. In Alignment mode, we compare an individual sense from a dataset we are importing to the senses of the term it is said to match in a language that is already in the Kamusi system. The task for the user is to select the correct match, if any. When a

threshold number of users agrees, we consider the relation confirmed. If results are ambiguous, we flag the item for manual review by a trusted user.

Future work will resolve how we treat items for which no match yet exists in our dataset. When the consensus is that an item does not have an existing sense match in Kamusi, we queue the pair for future consideration by a trusted user for the two languages being compared. A proper entry will need to be created for the corresponding language, which cannot be achieved through crowd microtasking.

3.7. Equivalence. Whether a bilingual pair is confirmed through the Translation mode or Alignment mode, we wish to know the degree of equivalence between the terms. Concepts might be functionally identical, e.g. a male cow will have parallel referents across virtually every language group that raises cattle. Concepts might have substantial similarities, but also notable differences, such that English "bull" as in bluster might be matched to something that has a more malicious connotation in another language, but is nevertheless the closest translation equivalent. Or, terms might be created in one language in the dictionary purely for the purpose of explaining a concept indigenous to another language, such that an explanatory phrase created in Mongolian to encapsulate a "bull" as an optimistic investor. In Equivalence mode, we show a bilingual pair and ask if the terms are parallel, similar, or an explanation on one side of an idea indigenous to the other. When a term passes consensus, the equivalence relation is published to the entries on both sides.

Future work will create a "Difference" mode for users to provide an explanation of the divergence between concepts marked "similar". This mode will operate along the lines of Definitions, with users able to write their own explanation or vote on explanations provided by others. In addition, the Difference mode will be employed for items shown as synonyms within a language, to distinguish, e.g., the difference between a boat and a ship.

4. Reasoning for Method Design

The tools discussed in section 3 were devised around a set of principles, some tested and some subject to testing. The methods remain experimental, and will be adjusted as the results become manifest. Notably, the methods discussed herein are not suitable for eliciting terms for which we do not already have a defined concept in a reference language, though future games could be built around processes such as the Rapid Word Collection technique (Moe 2007). In this section, we lay out the logic of the current instantiation.

We begin by noting that "ground truth" does not exist for most lexicographic tasks, in the form of digitized data that is comprehensive and indisputable. Datasets have not been compiled with the rigor we seek for languages that have some documentation, and are not available to science for most. Where data exists, it can never be considered definitive – knowing that a place to hold miscreants captive is a jail does not exclude the possibility that it is a prison, a penitentiary, or in Australia a gaol. However, we can find answers

that contain truth, and we can determine when information is clearly wrong. Without getting mired in well-trod linguistic theory, language is about communicating between people. If one person uses a word and the other person understands what is meant, communication has occurred. If a group of people all understand the same thing from the same word in the same context, that term can be considered part of their language. Roughly stated, a language consists of the set of expressions that are generally mutually agreed upon by its speakers, with the possibility of variations among regions or groups. All English speakers will agree about the basic sense of "water", however differently pronounced, for example, while Americans will be content to let South Africans have "robots" instead of "stoplights", and landlubbers accept that they will adopt an existing specialized boating vocabulary should they ever wish to sail.

Linguistic "truth" is pliable, but we posit that language groups can reach agreements that generally work. We deduce consensus when a number of participants repeat information, or when information provided by one is agreed acceptable by others. In many cases, the iterative process of seeking and validating information from a language's speakers may result over time in data that is qualitatively better than can be gathered by a solo lexicographer

In modes such as Translation, no user sees what has been proposed by others. We seek a participant's first reaction, and if other people respond the same way, we can be confident that a common basis for communication has been located. That is, if we ask for a term for a timepiece that is worn on the wrist, and most people answer "watch", then we have achieved a truth that many speakers of the language share that understanding of that term.

Modes such as Synonyms and Examples limit people to either/or answers. Questions are based on review of imported data or submissions from other users that can be resolved as right or wrong. Here, our premise is that people can usually agree about whether a data element is factual; "wristwatch" is a legitimate alternate term for a timepiece that is worn on the wrist, while "wratch" is not. If voting results are ambiguous, though, we acquire interesting information: we are alerted to the existence of conflicting opinions (Chklovski and Mihalcea 2003, Zhou and Li 2010, Aroyo and Welty 2013). These conflicts reveal non-binary facets of language, such as regional variation. When Boolean questions yield ambiguous answers, the results are queued for human review.

Definitions mode and the upcoming Difference mode rely on a mix of competitiveness and ego. When played as a game, the person who writes a winning definition receives many points, while people who vote for a winning definition each receive a single point. There is therefore substantial incentive to write a good definition if one has not already been proposed, but a disincentive to waste time challenging quality work. Moreover, the writer of a definition is given authorship credit online, but any definition can be challenged; readers thus have the incentive to submit improved definitions and thereby receive authorship credit. Of course, many participants will help improve definitions for the purely altruistic reason of contributing to the knowledge base, as often happens with Wikipedia, but we anticipate that motivating people to contribute for fun and recognition is a faster route to data collection that will nonetheless tend toward excellence over time.

The number of people required for an item to pass a consensus threshold is not firm. For starters, different modes have different considerations; more people must agree on the translation of an English noun than must verify that the translation is also a noun in their language. Within a mode, we can more quickly accept agreement that information is bad than that it is good, since "good" rejected data should someday reemerge, whereas publishing bad information undermines the confidence people can place in our results. It is impossible to say mathematically when validity is established; if one person proposes something, we know it is not yet valid, but what if two other people agree with the answer? Seven? 45? If five agree and four disagree, we should be less confident than seven agreeing and two disagreeing, but how much less, and how long do we keep asking? If four agree and five disagree, do we eliminate the answer? We are still tweaking the algorithm as we learn from experience, with the premise that answers that achieve unanimous support require a lower threshold than answers that generate conflict, and answers that generate too much noise should be passed to a trusted user for manual resolution. Thresholds may be adjusted for languages with few participants, if we can determine the credibility of the individual players, lest game data never acquire enough votes to enter the system. Users also gain trust by consistently being on the consensus side, with their votes acquiring increasing weight. Finally, the threshold can be overridden for a particular trusted user, such as a field linguist, who uses the system as a controlled elicitation tool, rather than as a game.

5. Future Work

Additional modes are planned for future crowdsourced microtasking. Among the tasks will be geotagging of usage sightings to build vocabulary clouds for the study of dialect, user contributions of pronunciations geotagged to their location of origin, and cataloguing of user-uploaded photos to illustrate terms and confirm labels across languages. For the present, the chief task is refining the processes based on user experience, and rolling out new languages in a manner that does not overwhelm our processing capacity, at the level of both management and machine.

6. Conclusions.

This paper has spelled out the design and reasoning behind the crowdsourcing features the Kamusi Project has introduce to acquire lexicographical data across numerous languages. Elsewhere we have discussed the underlying structure of the multilingual data, the multiplicity of approaches to data collection, and the specific use of gamification for harvesting social energy in the cause of lexicography. Our underlying premises are that much language knowledge can be affirmed by the consensus of its speakers about what elements achieve a communicative function, and that we can find that consensus by asking well-crafted questions to multiple speakers as targeted microtasks. However, we also find that some crowd results reveal ambiguity; this ambiguity produces valuable information about which data should be channeled for further analysis by crowd methods or designated language specialists. Our distributed approach is intended to elicit data for languages that would not otherwise receive attention from lexicographers, and also to elicit more extensive data than many existing lexicographic projects are able to acquire through expensive and time-consuming processes of data collection in the field or

through techniques such as corpus analysis. Crowdsourcing is not meant to be the exclusive method by which data is gathered and validated within the Kamusi system, but it provides an essential path to a wide range of lexical data that would not be available within a realistic timeframe via other methods.

References:

Aroyo, Lora, and Chris Welty. 2013. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *WebSci'13*, May 1-5, Paris, France.

Benjamin, Martin. 2013. *How We Chose a Priority List for Dictionary Entries*. http://kamusi.org/priority-list

Benjamin, Martin. 2014 a. Collaboration in the Production of a Massively Multilingual Lexicon. In *Conference Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland. http://www.lrec-conf.org/proceedings/lrec2014

Benjamin, Martin. 2014 b. Elephant Beer and Shinto Gates: Managing Similar Concepts in a Multilingual Database. In *Conference Proceedings of the 7th International Global WordNet Conference*, Tartu, Estonia.

Benjamin, Martin. 2014 c. *Molecular Lexicography: A lexical data model for Human Language Technology*. http://kamusi.org/molecular_lexicography

Benjamin, Martin and Paula Radetzky. 2014 a. Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification. *Presented at the 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, May 26-31.

Benjamin, Martin, and Paula Radetsky. 2014 b. Small Languages, Big Data: Multilingual Computational Tools and Techniques for the Lexicography of Endangered Languages. *Presented at ComputEL workshop, 52nd Meeting of the Association for Computational Linguistics*, Baltimore, June 22-27.

Brabham, Daren. 2011. Crowdsourcing: A model for leveraging online communities. in *The Participatory Cultures Handbook*, eds. Aaron Delwiche and Jennifer Jacobs Henderson, 120 -129. New York: Routledge.

Chklovski, Timothy, and Rada Mihalcea. 2003. Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. Proceedings of *RANLP*.

Hamari, Juho and Jonna Koivisto. 2013. Social Motivations to Use Gamification: An Empirical Study of Gamifying Exercise. In *ECIS 2013 Completed Research*. Paper 105.

Hirth, Matthias, Tobias Hoßfeld, and Phuoc Tran-Gia. 2013. Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. *Mathematical and Computer Modelling 57* (2013): 2918-2932

Hsueh, Pei-Yun, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing*, pages 27–35, Boulder, Colorado

Kittur, Aniket, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings, CHI 2008*, Florence, Italy.

McGonigal, Jane. 2011. *Reality Is Broken: Why games make us better and how they can change the world*. New York: Penguin.

Moe, Ronald. 2007. Dictionary Development Program. *SIL Forum for Language Fieldwork*, Linguistics, Language Technology

Saxton, Gregory D., Onook Oh, and Rajiv Kishore. 2013. Rules of Crowdsourcing: Models, Issues, and Systems of Control. *Information Systems Management* 30 (1): 2-20.

Schenk, Eric, and Claude Guittard. 2009. Crowdsourcing: What can be outsourced to the crowd, and Why? https://halshs.archives-ouvertes.fr/halshs-00439256v1

Vaish, Rajan, et al. 2014. Twitch Crowdsourcing: Crowd Contributions in Short Bursts of Time. *CHI 2014 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3645-3654.

Werbach, Kevin, and Dan Hunter. 2015. *The Gamification Toolkit: Dynamics, Mechanics, and Components for the Win*. E-book, Wharton School, University of Pennsylvania.

Zhou, Zhi-Hua, and Ming Li. 2010. Semi-supervised learning by disagreement. In *Knowledge Information Systems 24*:415–439